脉冲波形甄别中的最佳特征子集选择与降维研究

丁廷梦[1] 蒋宇航[1] 杨录英[1] 蒋小菲[1]

1(贵州大学 大数据与信息工程学院 贵阳 550025)

摘要 随着机器学习在中子-伽马（n-γ）甄别中的广泛应用，脉冲波形甄别中的特征子集选择成为一个值得关注的问题。经验方法、Random Forest 分类和 Logistic 回归特征选择算法较为全面地完善了特征子集选择方法，核主成分分析（KPCA）则将特征子集进一步降维。实验结果表明，特征选择算法在微弱的核信号中表现不佳，错误率均达 30%以上。经验方法中的特征子集选取范围则至关重要，特征子集“1-62”的错误率达到 49.096%，远高于来自脉冲尾部的特征子集约 1%的错误率。最优特征子集与尾积分对应的采样点不完全重合，但差异不大，尾积分对应的采样点可近似为最优特征子集。通过研究目前具有代表性的 Random Forest 分类、Logistic 回归等特征选择算法和细致的经验方法，论文结果具有普适性，为特征子集的选择提供了进一步的理论支持。

关键词：脉冲波形甄别、特征子集、特征选择、降维

Study on optimal feature subset selection and dimensionality reduction in pulse shape discrimination

DING Tingmeng[1] JIANG Yuhang[1] YANG Luying[1] JIANG Xiaofei[1]

1(*Big Data and Information Engineering, GuiZhou University, Guiyang 550025, China*)

Abstract **[Background]**:With the widespread application of machine learning in neutron-gamma (n-γ) discrimination, the selection of feature subsets in pulse waveform discrimination has become a notable issue. **[Purpose]**: By investigating representative feature selection algorithms such as Random Forest classification and Logistic regression, as well as detailed empirical methods, the results of this paper are universally applicable, providing further theoretical support for the selection of feature subsets. **[Methods]**: Empirical methods, Random Forest classification, and Logistic regression feature selection algorithms have comprehensively improved the methods of feature subset selection, while Kernel Principal Component Analysis (KPCA) further reduces the dimensionality of feature subsets. **[Results]**: Experimental results indicate that feature selection algorithms perform poorly in weak nuclear signals, with error rates exceeding 30%. The selection range of feature subsets in empirical methods is crucial, with error rates reaching 49.096% for feature subset "1-62", significantly higher than the approximately 1% error rate from features originating from the pulse tail. **[Conclusions]**: The optimal feature subset does not entirely overlap with the sampling points corresponding to the tail integral, but the difference is minor, suggesting that the sampling points corresponding to the tail integral can be approximated as the optimal feature subset.

Key words Pulse shape discrimination, Feature subset, Feature selection, Dimensionality reduction

---

国家自然科学基金(No .12205062) 、贵州省科技计划项目(黔科合 LH 字[2017]7225 号)资助

第一作者：丁廷梦，男，1998 年出生，2021 年毕业于东北大学理学院，现为硕士研究生，研究领域为中子伽马甄别

通信作者：蒋小菲，E-mail：514651931@qq.com

在多数中子场中，γ 射线总是伴随着中子而产生。目前常用的闪烁体探测器对中子和 γ 射线都很敏感，所以中子-伽马(n-γ)甄别是中子探测中的关键问题之一。

电荷比较法(CCM)是一种经典的 n-γ 甄别方法，该方法计算简单，且在高能域甄别效果极佳，其在脉冲甄别固件中被广泛应用。但是，CCM 无法甄别能量较低的脉冲。随着技术的发展，机器学习方法凭借其在分类和回归问题中的突出表现，已经被许多研究人员应用到 n-γ 甄别领域[1][2][3]。

机器学习中的无监督学习算法根据数据在特征空间的分布进行聚类，不依赖预先标记的样本，且具备一定的识别异常脉冲事件的能力，适用于中子伽马混合场。无监督学习中常用的高斯混合模型（GMM）在 n-γ 甄别中表现良好[4][5]，但是对高维度的数据直接进行 GMM 聚类时存在“维度灾难”。一个核脉冲信号包含上百个采样点，这样的高维数据直接用于 GMM 聚类时，不仅计算量大，而且聚类结果也很差。为了降低数据的维度，目前常用的方法是以经验选择脉冲差异较大的采样点作为特征子集，然后使用降维算法得到新的特征以实现降维。Liu et.al. (2023)使用脉冲尾部 14 个采样点作为特征子集，然后使用核主成分分析（KPCA）提取出 4 个新特征[6][7]。胡万平(2024)使用 KPCA-MPA-ELM 模型提升了 n- γ 甄别精度[8]，其中特征子集为脉冲尾部 80 个采样点，KPCA 用于对特征进行降维。

经验方法挑选的特征子集往往波动较大，特征子集中采样点的具体选取依据仍需要进一步研究。此外，特征选择算法也是重要的获得特征子集的方法[9][10]，但是在 n-γ 甄别中对特征选择算法的研究鲜少。如果能将特征选择算法也用于脉冲特征，能够进一步完善 n-γ 甄别中获得特征子集的方法。特征子集内一般含十余个特征，KPCA 可以对特征子集降维，进一步提取特征。

数据和特征决定了机器学习的上限，而模型和算法只是逼近这个上限。选择特征子集和维度降低有助于从原始数据中获取更多的信息，继而得到更准确的结果。本文首先使用三种方法得到特征子集：选择脉冲差异较大采样点的经验方法，随机森林(Random Forest)分类[11]特征选择算法和 Logistic 回归[12]特征选择算法。然后使用 KPCA 对特征子集进一步降维，最后通过更精细的经验特征子集得到最佳特征子集的选择范围。

1 方法与原理

在机器学习中，高维数据不适合直接作为输入。为了尽可能降低特征的维度，本文通过经验方法挑选脉冲差异较大的采样点、使用 Random Forest 分类和 Logistic 回归等特征选择算法得到特征子集，特征子集再降维以得到数个最佳低维特征(图 1)。

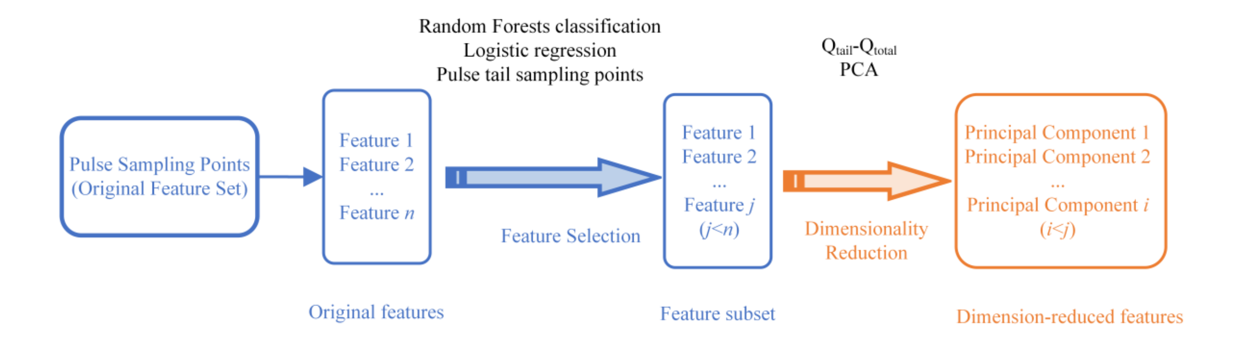


图 1 核脉冲信号中特征选择与降维示意图

Fig.1 Flowchart of impulse signal feature selection and dimensionality reduction

1.1 构建特征子集

面对高维数据，需要尽量去除冗余项，从原始脉冲采样点中挑选一部分特征组成特征子集。特征子集可以通过经验挑选脉冲差异较大的采样点得到，还可以通过 Random Forest 分类和 Logistic 回归等特征选择算法得到。

脉冲尾部是中子和伽马射线两类脉冲间差异最大的部分，特征子集内的特征应当来自于脉冲尾部的采样点，但采样点的多寡并不统一。构建特征子集的常规做法是以经验挑选脉冲差异较大的采样点，即脉冲尾部的数十个点，当然该方法依赖于“经验”与估计。

特征选择算法通过去除不相关、冗余或嘈杂的特征，从原始特征中选择小部分特征进行降维，Random Forest 分类和 Logistic 回归都属于特征选择方法。随机森林是一种包含多棵决策树的分类器算法模型，而每棵决策树由根节点、内部节点和叶节点组成。叶子节点为分类结果，根节点和内部节点为决策依据。Logistic 回归是一种二分类算法，通过多个自变量的线性组合来预测分类变量的概率。

为了增加 Random Forest 特征选择算法结果的可靠性，我们逐步增大原始特征的大小，原始特征分别取脉冲尾部 34 个采样点、脉冲非基线部分 62 个采样点以及包含部分基线的 120 个采样点。通过特征子集内包含的脉冲与脉冲尾部采样点之间的差异，我们能够评估 Random Forest 分类特征选择算法对冗余项的排除能力。Random Forest 分类和 Logistic 回归从原始特征中会分别选择出一个特征子集，如果特征选择算法的可靠性高，Random Forest 分类和 Logistic 回归所得的特征子集应当保持一致；如果核脉冲采样点中有重要性明显高的特征，那基线不应对特征选择结果产生影响，将基线采样点也纳入原始特征可以评估特征选择算法的稳定性。

1.2 降维

核主成分分析（KPCA）是一种基本的特征提取方法，该方法将高维数据映射到低维正交特征上，这些重新构造的特征被称为主成分。特征子集内的特征数依旧较高，KPCA 将特

征子集内的特征映射为新的主成分从而实现降维，此时只需要数个主成分即可得到较高的累计方差。

1.3 Q_{tail} 和 Q_{total}

如图 2 所示，Q_{tail} 和 Q_{total} 分别是脉冲的电荷总积分和尾积分，CCM 正是以这二者比值作为甄别因子。Q_{tail} 和 Q_{total} 可视为从脉冲非基线采样点提取的两个独立特征，Q_{tail} 和 Q_{total} 可以作为 GMM 聚类的特征获得比 CCM 更好的分类结果。

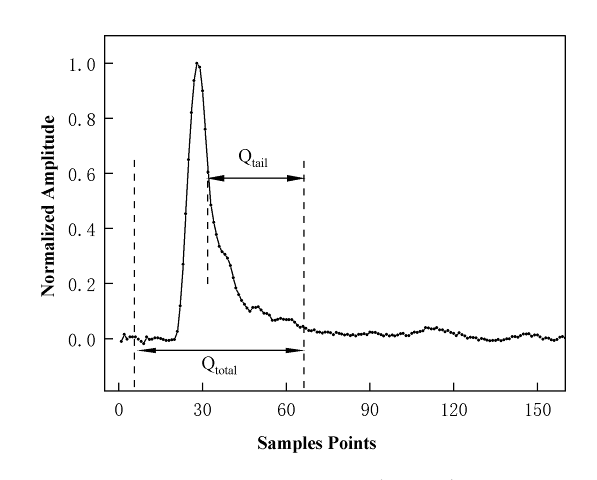


图 2 Q_{tail} 和 Q_{total} 的选取示意图
Fig.2 Diagram of the tail integral Q_{tail} and the total integral Q_{total}

1.4 GMM 聚类

高斯混合模型(Gaussian Mixture Model，GMM)是一种概率模型，用于描述由多个高斯分布组成的数据集。对于每个分量高斯，其概率密度函数为：

$$f(\boldsymbol{x}|\boldsymbol{\mu},\Sigma)=\frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}e^{-\frac{1}{2}}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \tag{1.1}$$

其中 n 是数据的维度，$\boldsymbol{\mu}$是 n 维均值向量，Σ是 $n\times n$ 的协方差矩阵，显然高斯分布由$\boldsymbol{\mu}$和Σ确定。初始的脉冲具有 248 个特征，如此高的维度会存在“维度灾难”。为减少脉冲的特征数量，需要先进行特征提取或者选择。

忽略脉冲堆积，在 n-γ 甄别中，该模型只存在 neutrons 和 gamma rays 两个成分。对于两个混合成分的高斯混合分布，其概率密度为：

$$f_M(\boldsymbol{x})=\sum_{i=1}^{2}\alpha_i\cdot f(\boldsymbol{x}|\boldsymbol{\mu}_i,\Sigma_i) \tag{1.2}$$

α_i 称为“混合系数”，它为选择第 i 个高斯混合成分的概率。其中$\alpha_i>0$ 且$\sum_{i=1}^{2}\alpha_i=1$。

模型参数α_i、$\boldsymbol{\mu}_i$ 和Σ_i 需要通过 EM 算法迭代优化求解。EM 算法每一步迭代包括两步：E 步，根据当前的参数估计隐变量的期望；M 步，利用 E 步的计算结果，根据最大似然估计更新模型参数。

当输入 GMM 聚类的特征不同时，聚类结果也会随之变化。本文旨在探究最优的特征子集获取方法，通过比较不同的特征子集进行 GMM 聚类后的结果可以评估不同特征子集的优劣。

2 结果和讨论

本次实验的流程图如**图 3** 所示，中子源为 ^{241}Am-Be 源，探测器是有机液体闪烁体探测器 EJ-301，数字化仪为 DT5730B。探测器采集到的是电流脉冲，脉冲经过数字化后得到原始数据，原始数据经过平滑滤波、归一化和基线恢复等预处理步骤后存储于计算机中[13]。预处理后的 60000 个脉冲分为两部分，其中 30000 个脉冲用于进行 GMM 聚类，以得到一个可靠的训练集；另外 30000 个脉冲用于测试，以比较不同算法的性能差异。

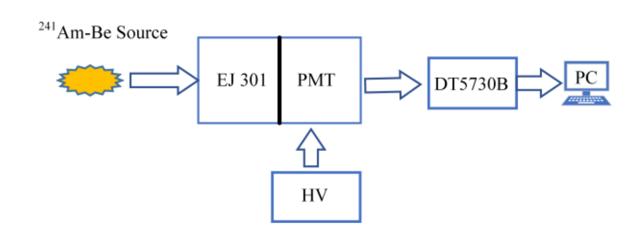


图 3 实验流程图
Fig.3 Experimental flowchart

2.1 特征子集构建

脉冲的尾部积分对应的采样点有 34 个，脉冲非基线部分包括 62 个采样点。中子伽马两类脉冲差异最大的部分是尾积分对应的 34 个采样点，特征子集可以通过经验挑选脉冲差异较大的采样点得到。

除了经验方法选取特征子集，我们也使用特征选择算法以完善脉冲形状甄别中的特征子集获取方法。Random Forest 分类是特征工程中获得特征子集重要方法，我们采用 5 折交叉验证，选定子集大小范围从 1 到 13。

Random Forest 分类和 Logistic 回归依赖于先验知识，我们取 GMM 聚类结果中概率大于 99%的脉冲作为训练集，依托该训练集寻找特征子集。以 Q_{tail} 和 Q_{total} 作为特征的 GMM 聚类在 100-2100 keV 内的分类结果与经典的 CCM 保持一致，在 0-100 keV 内的脉冲分类正确率比 CCM 高 5.52%。GMM 聚类的结果时概率值，排除低概率事件(分类概率<99%)后，剩余脉冲可构成一个大小为 26261 的训练集。

Random Forest 分类从脉冲尾部的 34 个采样点挑选特征子集时，所得特征子集仅包含两个特征，但是其重要性评分分别为 0.091 和 0.087，特征重要性很低，其用作聚类后的结果极差。

脉冲尾部是中子和伽马射线两类脉冲间差异最大的部分，最优特征子集内多数特征应当来自于脉冲尾部的采样点。为了评估特征选择方法的可靠性，我们将原始特征扩大至非基线部分的 62 个采样点，通过最优子集内包含的脉冲与脉冲尾部采样点之间的差异，评估该方法对冗余项的排除能力。**图 4** 展示了不同特征子集大小下的性能指标，包括均方根误差(RMSE, Root Mean Square Error)、确定系数(R-squared)和平均绝对误差(MAE, Mean Absolute Error)，这些指标用于衡量模型在不同特征子集大小下的预测准确度和稳定性。

图 4(a)和**图 4**(b)分别是 R-squared，RMSE 和 MAE 随子集计数变化的柱状图和折线图。随着特征子集内的特征数量从 1 到 13 依次递增时，R-squared 呈现上升趋势，RMSE 和 MAE 则是下降趋势。从特征计数大于 7 以后，R-squared，RMSE 和 MAE 变化率陡增。在特征计

数大于 10 后，R-squared 增长变缓慢，同时 RMSE 和 MAE 的下降变缓慢。从折线图中不难发现，特征计数大于 10 后三条折线的变化都趋缓。综合柱状图和折线图的结果，随机森林从 62 个特征中选择的最佳特征子集包含 10 个特征，这 10 个特征分别是第 23、22、24、6、16、9、32、21、23、35、15、48 个采样点。脉冲尾部是中子和伽马射线两类脉冲间差异最大的部分，最优特征子集内多数特征应当来自于脉冲尾部的采样点，但是该子集与脉冲下降沿采样点重合度不高，仅有 3 个采样点来自于脉冲尾部。

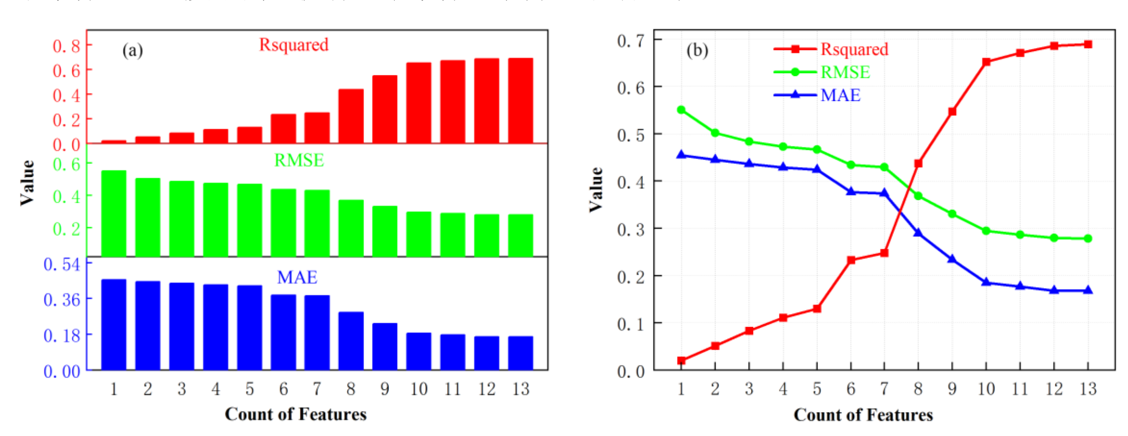


图 4 性能指标随着特征子集大小变化的柱状图和折线图(62 个采样点作为原始特征)

Fig.4 Histograms and line plots of changes in performance metrics with feature subset size (62 sample points)

因为基线调零后，脉冲的基线存在波动，不完全为 0。为了探究脉冲基线对特征选择算法的影响，基线的部分采样点也需要被考虑到特征选择算法中。当原始特征数为 120 时，特征选择的结果如**图 5** 所示，**图 5** (a)和**图 5** (b)分别是 R-squared，RMSE 和 MAE 随子集计数变化的柱状图和折线图。在特征计数大于 6 后，R-squared 增长变缓慢，同时 RMSE 和 MAE 的下降变缓慢。在 R-squared，RMSE 和 MAE 随子集计数变化的折线图中，特征计数大于 6 后三条折线的变化都趋缓。综合柱状图和折线图的结果，随机森林从 120 个特征中选择的最佳特征子集包含 6 个特征，这与从 62 个采样点中选择的特征子集不同。核信号是非常微弱的信号，我们无法彻底去除掉所有噪声，基线不可能完全为 0。使用 Random Forest 分类算法进行特征选择时，基线对特征选择的结果存在很大干扰，特征选择算法的抗干扰能力较差。

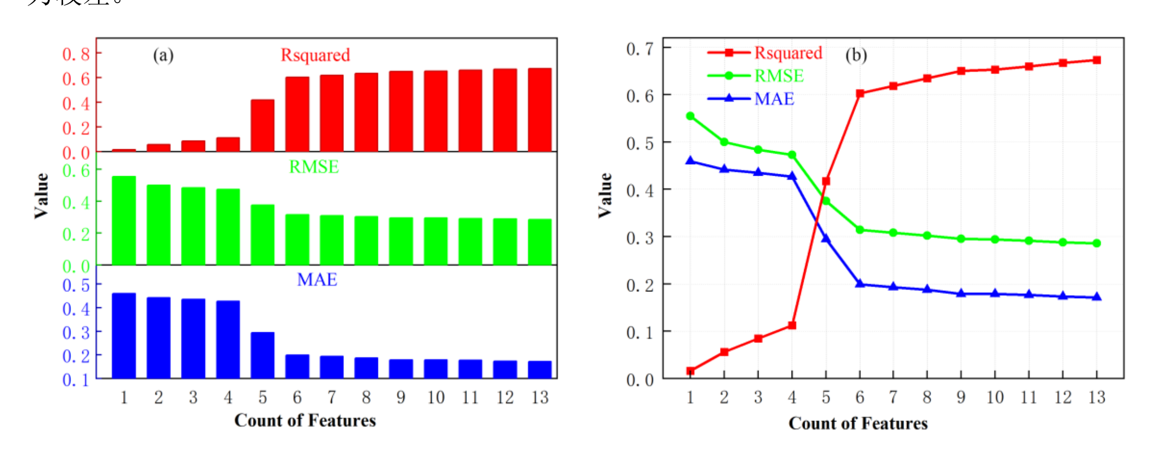


图 5 性能指标随着特征子集大小变化的柱状图和折线图(120 个采样点作为原始特征)

Fig.5 Histograms and line plots of changes in performance metrics with feature subset size (120 sample points)

Random Forest 分类和 Logistic 回归都是重要的特征选择算法，除了利用 Random Forest 分类模型进行特征选择，我们还可以利用 leaps 中的 Logistic 回归模型从 62 个尾部采样点中选择特征子集。Logistic 回归模型拟和的效果以四个参数评估，这四个参数分别是：残差平方和(RSS ,Residual Sum of Squares)，调整后的决定系数(Adjusted R2, Adjusted R-Squared)，Mallow's Cp (CP)和贝叶斯信息准则(BIC ,Bayesian Information Criterion)。

图 6 展示了在 Logistic 回归模型中，不同大小特征子集对应的性能指标。特征子集内的特征数为 11 时，CP 和 BIC 都达到最小值，与此同时 Adjusted R2 达到最大值，RSS 在特征数为大于 11 之后不再呈现明显的下降趋势。根据该图可知 Logistic 回归模型选择的最优特征子集包含 11 个特征。Random Forest 分类和 Logistic 回归两种特征选择算法的结果不同，表明特征选择算法难以获得可靠的稳定的特征子集。

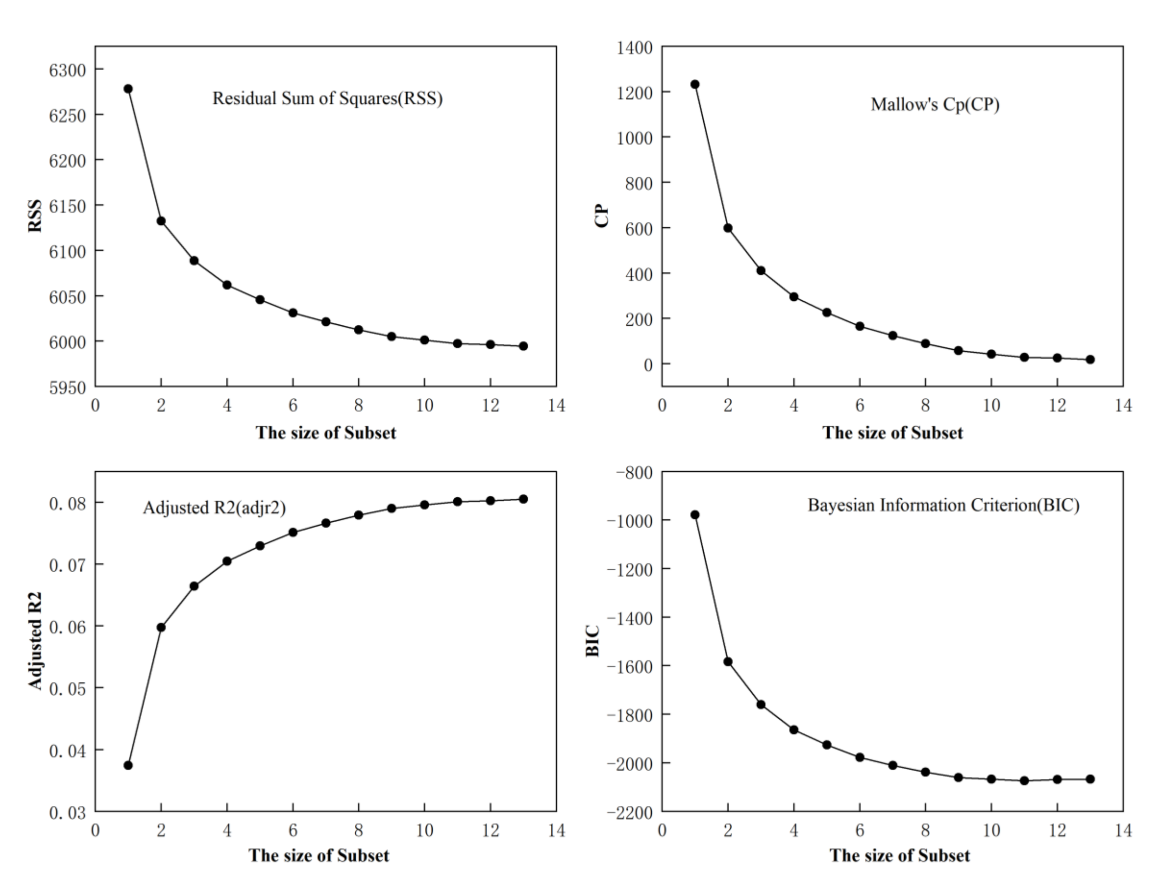


图 6 在 Logistic 回归模型中，不同特征子集大小下的性能指标
Fig.6 Performance metrics for different feature subset sizes in the regression model

2.2 最优特征子集

通过经验挑选、Random Forest 分类和 Logistic 回归等特征选择算法得到的特征子集并不相同。脉冲尾积分部分包括 34 个采样点（特征），Random Forest 分类模型挑选的特征子集包含 10 个特征，Logistic 回归模型挑选的特征子集包含 11 个特征。相比于脉冲的 248 个采样点，这些特征子集的大小以及大为缩小，我们可先以特征子集直接作为特征分析 GMM 聚类效果。为了比较不同方法甄别效果的差异，我们以不同的特征进行 GMM 聚类，并且按“特征数+GMM (*n*-features GMM)”命名不同的方法。10-features GMM、11-features GMM、

34-features GMM 和 62-features GMM 四种方法使用的特征分别对应于 Random Forest 分类模型挑选的特征子集、Logistic 回归模型挑选的特征子集、脉冲尾积分部分的 34 个采样点和脉冲非基线部分的 62 个采样点。

对于 EJ-301 探测器得到的脉冲数据，中子伽马在较低的能量域内完全混合，无法被甄别。要比较不同特征下聚类效果的差异，具有可靠标签的脉冲是必需的。CCM 是在中子伽马(n-γ)甄别中被广泛使用的一种经典方法，在较高的能量域(100-2100 KeV)内中子和伽马射线能够完全分离。CCM 在高能量域的甄别结果是可靠的，比较不同方法对 100-2100KeV 脉冲的分类结果，可以评估 n-γ 甄别的效果。

为了定量分析不同方法甄别结果之间的差异，我们将不同方法的结果进行两两得到甄别结果差异热力图如**图 7** 所示。不难发现，34 features GMM 与 CCM 甄别结果之间相差最小，只有 1.36%的差异，但是 10 features GMM（Random Forest 分类）、11 features GMM（Logistic 回归）和 62 features GMM（脉冲非基线部分的 62 个采样点）与 CCM 甄别结果差异都达到了 30%以上。此外，10 features GMM、11features GMM 和 62 features GMM 不仅甄别精度低，其甄别结果相互之间的差异也是巨大的。

在经验选择特征子集时，采样点的选择范围极为重要，62 features GMM 的错误率高达 30.06%，远高于 34 features GMM。一方面，62 features GMM 使用的特征不是脉冲差异最大的部分；另一方面，62 features GMM 特征维度依旧较高，存在“维度灾难”。

另一个很明显的事实是特征选择算法在脉冲波形甄别中的特征选择表现极差，同属于特征选择算法的 Random Forest 分类和 Logistic 回归之间的差异为 24.61%，且错误率均在 30%以上，特征选择结果不精确且不稳定。核脉冲是非常微弱的信号，易受噪声影响，单个采样点的波动较大，特征选择是从找出重要性最高的数个特征，但是核信号缺乏主导性的采样点，这导致特征选择在脉冲特征工程中表现不佳。

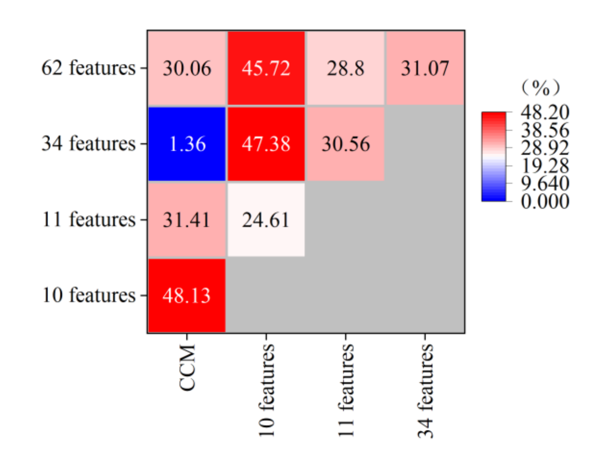


图 7 以不同特征子集进行 GMM 聚类时，分类结果的差异热力图
Fig.7 Differences in classification results with different feature subsets

Q_{tail} 和 Q_{total} 通过对采样点积分，降低了单个采样点的波动对甄别结果的影响。主成分分析计算新的正交特征，主成分的解释方差远高于其他特征，对甄别结果具有决定性作用。为了进一步探究最优特征子集的取值，在特征子集之后通过进一步降维，可以得到重要性更高

且更少的新特征。

2.3 KPCA 降维

脉冲非基线部分包括 62 个采样点，脉冲差异最大即脉冲尾部有 34 个采样点，这两个特征子集都是具有代表性的经验选择方法。

脉冲非基线部分包括 62 个采样点，使用 KPCA 对 62 个采样点进行降维后，我们取前三个主成分作为特征进行 GMM 聚类。在特征空间中，我们很难准确判断分类结果的准确性。**图 8** 是该方法的聚类结果在 Energy-PSD 图中的分布，其中正方形代表中子，圆形代表γ射线，该结果存在大量错误甄别。

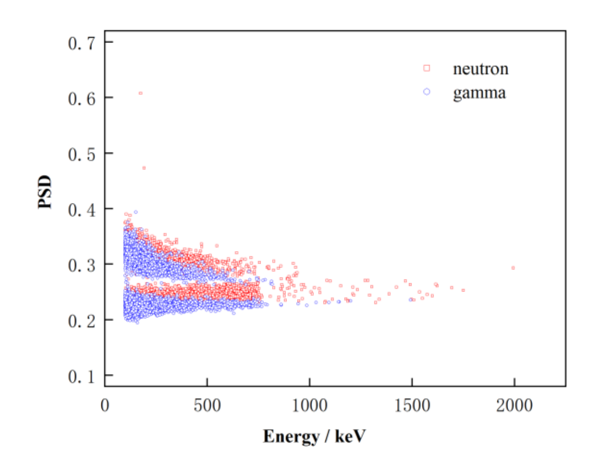


图 8 非基线采样点作为原始特征降维后，前三个主成分输入 GMM 聚类的结果
Fig.8 The non-baseline sampling points as raw features and the GMM clustering result of the first three principal components.

以脉冲尾部中随机 14 个采样点的 KPCA 结果为例，KPCA 的结果如**图 9** 所示，前三个主成分的解释方差分别为 64.65%、22.73%和 3.35%，图中前三个主成分累计方差(即三者之和)已经超过 90%，并且第一个主成分的占比极高。即使是从脉冲尾部随机取样，前三主成分依旧具有较高的解释方差。

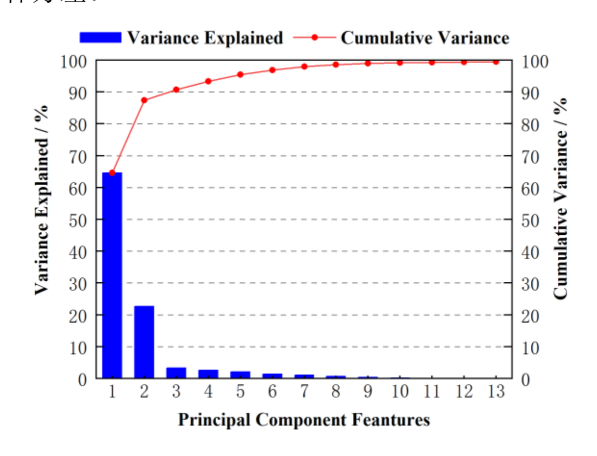


图 9 脉冲尾部中 14 个主成分特征的贡献率和累积贡献率
Fig.9 Contribution rate and cumulative contribution rate of 14 principal component features

为了进一步完善经验选取方法，我们从峰值位置(第 28 个采样点)作为起点，取 28 到 38 采样点作为第一个特征子集，每次增加三个采样点至第 62 个采样点，此时共有 9 个特征子集，大小从 10 到 34。此外，我们还增加了第 28 个采样点前的采样点以探究冗余项的影响。表 1 是经验选择方法的部分特征子集降维的结果，其中起止采样点是特征子集的选取范围；前三主成分累计方差反映了 KPCA 的结果，特征子集越小，前三主成分累计方差越高；误差则是在 100-2100 keV 能量域内，以三个主成分进行 GMM 聚类后的结果与标签对比后的错误率。特征子集“28-38”，“28-50”以及“28-62”都是来自于脉冲尾部的 34 个采样点，子集“25-62”和“1-62”则是引入了一部分脉冲非基线的采样点。不难发现，特征子集来自脉冲尾部采样点时，错误率均在 1%左右。特征子集“25-62”是错误率最低的，脉冲的尾积分起点位置是以最优的 CCM 甄别结果确定的，与最优 KPCA-GMM 聚类结果不完全相同，但差异不大。特征子集“1-62”的错误率达到 49.096%，该特征子集不论直接聚类还是 KPCA 后聚类，结果都很差。结合特征选择算法的结果，特征子集内的特征必须来自于脉冲差异最大部分的采样点，该部分与尾积分对应的采样点不完全重合，但差异不大，尾积分对应的采样点可近似为最优特征子集。

表 1 经验选择方法的部分特征子集降维结果
Table 1 The dimensionality reduction results of partial feature subsets selected by experience selection methods

起止采样点	28-38	28-50	28-62	25-62	1-62
前三主成分累计方差	0.9983	0.9890	0.9746	0.9707	0.9070
错误率	1.120%	1.019%	1.044%	0.929%	49.096%

3 结语

为了得到最佳特征子集，本文通过经验方法、Random Forest 分类特征选择算法和 Logistic 回归特征选择算法得到特征子集。经验挑选特征子集从 10 个采样点到 62 个采样点不等，Random Forest 分类得到的特征子集包括 10 个特征，Logistic 回归得到的特征子集包括 11 个特征。在经验选择特征子集时，采样点的选择范围极为重要，62 features GMM 的错误率高达 30.06%，远高于 34 features GMM。特征选择算法在脉冲波形甄别中的特征选择表现极差，Random Forest 分类和 Logistic 回归之间的差异为 24.61%，且错误率均在 30%以上，特征选择结果不精确且不稳定。

特征选择算法在特征选择中，面临着三个问题：首先，原始特征为脉冲非基线部分采样点（62 采样点）时，挑选的特征子集与尾部采样点重合度不高，这表明算法的特征选择精度低；其次，原始特征包含一部分基线部分采样点（120 采样点）时，特征子集与从脉冲非基线部分采样点中挑选的特征子集不同，这表明该方法稳定性差；最后，Random Forest 分类和 Logistic 回归两种特征选择算法的结果不同，这也表明特征选择算法难以获得可靠的稳定的特征子集。核信号是十分微弱的信号，单个采样点对脉冲甄别结果的影响有限，而且采样点数值波动大。因为没有起主导重要性的采样点，分类和回归的特征选择方法得到的结果不同，稳定性差，甚至于基线采样点也会对特征选择的结果有影响，从 62 个采样点得到的最优特征子集大小(10)与从 120 个采样点得到的最优特征子集大小(6)不同。

主成分分析计算新的正交特征，前三主成分的解释方差远高于其他特征。为了得到最优

特征子集，我们将经验选择得到的特征子集降维，进行了更细致的分析。特征子集来自于脉冲尾部时，错误率均在 1%左右。特征子集“1-62”的错误率达到 49.096%，该特征子集不论直接聚类还是 KPCA 后聚类，结果都很差。特征子集“25-62”是错误率最低的，这说明最优特征子集与尾积分对应的采样点不完全重合，但差异不大，尾积分对应的采样点可近似为最优特征子集。

设备

网络通信信号检测系统，1502195N；NI-1085 机箱、NI-5162 高速数字化仪、NI-7976R FPGA 模块。

作者贡献声明

丁廷梦负责数据处理、调查研究、可视化呈现、实验结果分析和论文写作；蒋小菲负责项目管理、提供资源、指导和审阅；蒋宇航负责调查研究；杨录英负责数据管理。

参考文献

1 Durbin M., Wonders M.A., Flaska M., *et al*. K-Nearest Neighbors regression for the discrimination of gamma rays and neutrons in organic scintillators，Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 2021, **987**:164826. DOI: 10.1016/j.nima.2020.164826

2 Arahmane H., Hamzaoui E.M., Maissa Y. B., *et al*. Neutron-gamma discrimination method based on blind source separation and machine learning. Nuclear Science and Techniques, 2021,**32**: 18. DOI: 10.1007/s41365-021-00850-w.

3 黄坤翔,张江梅,王嘉麒,等.基于 GAF-CNN 的 n/γ 甄别方法研究[J].原子能科学技术,2024,58(02):461-470. DOI: 10.7538/yzk.2023.youxian.0398.
HUANG Kunxiang, ZHANG Jiangmei, WANG Jiaqi, *et al*. Study on n/γ Discrimination Method Based on GAF-CNN[J]. Atomic Energy Science and Technology, 2024,**58**(02):461-470. DOI: 10.7538/yzk.2023.youxian.0398.

4 Andrew G., Qi C., Kaplan A.D., *et al*. Pulse pileup rejection methods using a two-component Gaussian Mixture Model for fast neutron detection with pulse shape discriminating scintillator[J]. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment,2021, **988**: 164905. DOI: 10.1016/j.nima.2020.164905.

5 Wang F P., Yang M H., Wang J Y., *et al*. A comparison of small-batch clustering and charge-comparison methods for n/γ discrimination using a liquid scintillation detector. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 2022,**1028**: 166379. DOI:10.1016/j.nima.2022.166379.

6 Liu L F, Shao H. Study on neutron-gamma discrimination method based on the KPCA-GMM-ANN[J]. Radiation Physics and Chemistry, 2023, **203**: 110602. DOI: 10.1016/ j.radphyschem.2022.110602.

7 Liu L F, Shao H. Study on neutron-gamma discrimination method based on the KPCA-GMM[J]. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, 2023, **1056**: 168604. DOI: 10.1016/j. nima. 2023.168604.

8 胡万平,张贵宇,张云龙,等.基于 KPCA-MPA-ELM 的 n/γ 甄别方法研究[J].核技术,2024,47(04):75-84. DOI: 10.11889/j.0253-3219.2024.hjs.47.040403

HU Wanping, ZHANG Guiyu, ZHANG Yunlong, *et al*. Neutron/gamma (n/γ) discrimination method based on KPCA-MPA-ELM[J]. NUCLEAR TECHNIQUES, 2024,**47**(04):75-84. DOI: 10.11889/j.0253-3219.2024.hjs.47.040403

9 张金区,凌毓,杜平,等.面向单脉冲信号分类的集成特征选择与评价[J].天文学报,2023,64(05):59-69. DOI: 10.15940/j.cnki.0001-5245.2023.05.006.

ZHANG Jin-qu, LING Yu, DU Ping, *et al*. Ensemble Feature Selection Method for Single Pulse Classification[J]. Acta Astronomica Sinica, 2023,**64**(05):59-69. DOI: 10.15940/j.cnki.0001-5245.2023.05.006.

10 Ishwaran H., Malley J. D. Synthetic learning machines[J]. Biodata Mining,2014,7. DOI: 10.1186/s13040-014-0028-y.

11 Vladimir S, Andy L, Christopher T, *et al*. A Classification and Regression Tool for Compound Classification and QSAR Modeling[J]. Journal of Chemical Information and Computer Sciences, 2003, **43** (6): 1947-1958. DOI: 10.1021/ci034160g

12 Nuttanan W, Kang Y Y, Zhang F Q. Random feature selection using random subspace logistic regression[J]. Expert Systems with Applications, 2023, **217**(119535): 0957-4174.DOI: 10.1016/j.eswa.2023.119535.

13 汪炫羲. 闪烁体探测器的 n-γ 脉冲波形甄别研究[D]. 贵州大学,2023. DOI: 10.27047/d.cnki.ggudu.2023.001514.

Wang X X. The n-γ pulse waveform discrimination of scintillator detectors[D]. GuiYang: GuiZhou University,2023. DOI: 10.27047/d.cnki.ggudu.2023.001514.